RESEARCH ARTICLE SUMMARY

DISEASE NETWORKS

Uncovering disease-disease relationships through the incomplete interactome

Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, Albert-László Barabási*

INTRODUCTION: A disease is rarely a straightforward consequence of an abnormality in a single gene, but rather reflects the interplay of multiple molecular processes. The relationships among these processes are encoded in the interactome, a network that integrates all physical interactions within a cell, from protein-protein to regulatory protein–DNA and metabolic interactions. The documented propensity of disease-associated proteins to interact with each other suggests that they tend to cluster in the same neighborhood of the interactome, forming a disease module, a connected subgraph that contains all molecular determinants of a disease. The accurate identification of the corresponding disease module represents the first step toward a systematic understanding of the molecular mechanisms underlying a complex disease. Here, we present a network-based framework to identify the location of disease modules within the interactome and use the overlap between the modules to predict disease-disease relationships.

RATIONALE: Despite impressive advances in high-throughput interactome mapping and disease gene identification, both the interactome and our knowledge of disease-associated genes remain incomplete. This incompleteness prompts us to ask to what extent the current data are sufficient to map out the disease modules, the first step toward an integrated approach toward human disease. To make progress, we must formulate mathematically the impact of network incompleteness on the identifiability of disease modules, quantifying the predictive power and the limitations of the current interactome.

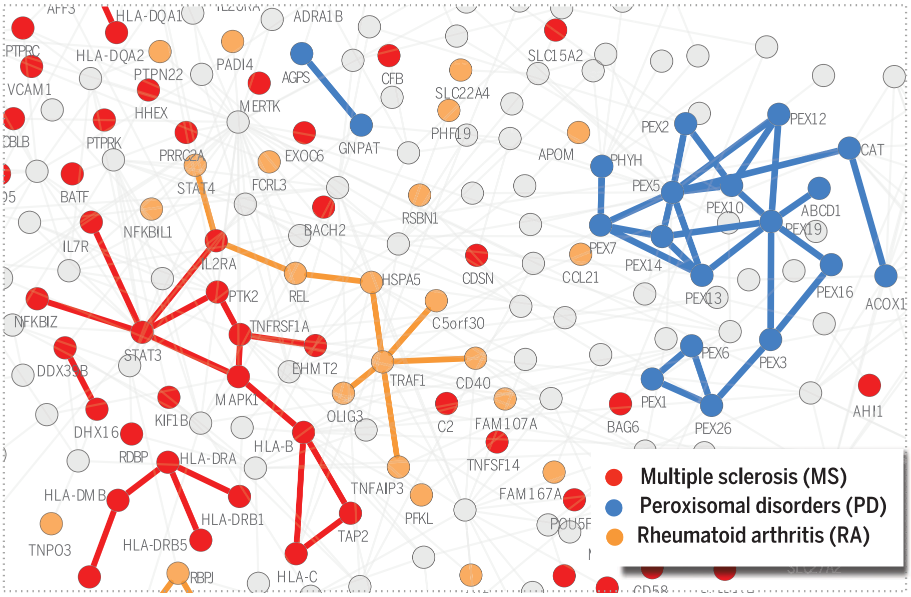
RESULTS: Using the tools of network science, we show that we can only uncover disease modules for diseases whose number of associated genes exceeds a critical threshold determined by the network incompleteness. We find that disease proteins associated with 226 diseases are clustered in the same network neighborhood, displaying a statistically significant tendency to form identifiable disease modules. The higher the degree of agglomeration of the disease proteins within the interactome, the higher the biological and functional similarity of the corresponding genes. These findings indicate that many local neighborhoods of the interactome represent the observable part of the true, larger and denser disease modules.

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.1257601

If two disease modules overlap, local perturbations causing one disease can disrupt pathways of the other disease module as well, resulting in shared clinical and pathobiological characteristics. To test this hypothesis, we measure the network-based separation of each disease pair, observing a direct relation between the pathobiological similarity of diseases and their relative distance in the interactome. We find that disease pairs with overlapping disease modules display significant molecular similarity, elevated coexpression of their associated genes, and similar symptoms and high comorbidity. At the same time, nonoverlapping disease pairs lack any detectable pathobiological relationships. The proposed network-based distance allows us to predict the pathobiological relationship even for diseases that do not share genes.

CONCLUSION: Despite its incompleteness, the interactome has reached sufficient coverage to allow the systematic investigation of disease mechanisms and to help uncover the molecular origins of the pathobiological relationships between diseases. The introduced network-based framework can be extended to address numerous questions at the forefront of network medicine, from interpreting genome-wide association study data to drug target identification and repurposing. ■



Diseases within the interactome. The interactome collects all physical interactions between a cell's molecular components. Proteins associated with the same disease form connected subgraphs, called disease modules, shown for multiple sclerosis (MS), peroxisomal disorders (PD), and rheumatoid arthritis (RA). Disease pairs with overlapping modules (MS and RA) have some phenotypic similarities and high comorbidity. Non-overlapping diseases, like MS and PD, lack detectable clinical relationships.

The list of author affiliations is available in the full article online.
*Corresponding author. E-mail: alb@neu.edu

Cite this article as J. Menche *et al.*, *Science* **347**, 1257601 (2015). DOI: 10.1126/science.1257601

RESEARCH ARTICLE

DISEASE NETWORKS

Uncovering disease-disease relationships through the incomplete interactome

Jörg Menche,[1,2,3] Amitabh Sharma,[1,2] Maksim Kitsak,[1,2] Susan Dina Ghiassian,[1,2] Marc Vidal,[2,4] Joseph Loscalzo,[5] Albert-László Barabási[1,2,3,5]*

According to the disease module hypothesis, the cellular components associated with a disease segregate in the same neighborhood of the human interactome, the map of biologically relevant molecular interactions. Yet, given the incompleteness of the interactome and the limited knowledge of disease-associated genes, it is not obvious if the available data have sufficient coverage to map out modules associated with each disease. Here we derive mathematical conditions for the identifiability of disease modules and show that the network-based location of each disease module determines its pathobiological relationship to other diseases. For example, diseases with overlapping network modules show significant coexpression patterns, symptom similarity, and comorbidity, whereas diseases residing in separated network neighborhoods are phenotypically distinct. These tools represent an interactome-based platform to predict molecular commonalities between phenotypically related diseases, even if they do not share primary disease genes.

Identifying sequence variations associated with specific phenotypes represents only the first step of a systematic program toward understanding human disease. Indeed, most phenotypes reflect the interplay of multiple molecular components that interact with each other (*1*–*6*), many of which do not carry disease-associated variations. Hence, we must view disease-associated mutations in the context of the human interactome, a comprehensive map of all biologically relevant molecular interactions (*6*–*12*).

Yet, the predictive power of the current network-based approaches to human disease is limited by several conceptual and methodological issues. First, high-throughput methods cover less than 20% of all potential pairwise protein interactions in the human cell (*11*–*16*), which means that we seek to discover disease mechanisms relying on interactome maps that are 80% incomplete. Second, the genetic roots of a disease are traditionally captured by the list of disease genes whose mutations have a causal effect on the respective phenotype. The disease proteins (the products of disease genes) are not scattered randomly in the interactome, but tend to interact with each other, forming one or several connected subgraphs that we call the disease module (Fig. 1A). This agglomeration of disease proteins is supported by a range of biological and empirical evidence (*7*, *17*, *18*) and has fueled the development of numerous tools to identify new disease genes and prioritize pathways for disease relevance (*8*, *9*, *19*–*28*). Despite its frequent use, however, the disease module hypothesis lacks a solid mathematical basis. Third, the relationships between distinct phenotypes are currently uncovered by identifying shared components like disease genes, single-nucleotide polymorphisms (SNPs), pathways, or differentially expressed genes involved in both diseases. This has resulted in the construction of "disease networks," unveiling the common genetic origins of many disease pairs (*7*, *29*). Yet, shared genes offer only limited information about the relationship between two diseases. Indeed, mechanistic insights are often carried by the molecular networks through which the gene products associated with the two diseases interact with each other.

The fragmentation of disease modules

We started by compiling 141,296 physical interactions between 13,460 proteins experimentally documented in human cells, including protein-protein and regulatory interactions, metabolic pathway interactions, and kinase-substrate interactions [Fig. 1; see also figs. S1 and S2 and supplementary materials (SM) section 1 for a detailed discussion], representing a blueprint of the human interactome (Fig. 1D). We also compiled a corpus of all 299 diseases defined by the Medical Subject Headings (MeSH) ontology that have at least 20 associated genes in the current Online Mendelian Inheritance in Man (OMIM) and genome-wide association study (GWAS) databases (*30*, *31*), involving 2436 disease-associated proteins (Fig. 1, B and C, and SM section 1).

Despite the best curation efforts, both the interactome and the disease gene list remain incomplete (*6*, *11*–*16*) and biased toward much-studied disease genes and disease mechanisms (*32*, *33*). The consequences of this incompleteness are illustrated by multiple sclerosis: Of the 69 genes associated with the disease, only 11 disease proteins form a connected subgraph (observable module, Fig. 1D); the remaining 58 proteins appear to be distributed randomly in the interactome. This pattern holds for all 299 diseases, their observable modules comprising on average only 20% of the respective disease genes (Fig. 1C). Several factors contribute to this fragmentation (Fig. 1A), the main one being data incompleteness: Missing links leave many disease proteins isolated from their disease module (Fig. 1A).

In percolation theory, if only a p fraction of links is available, a connected subgraph (disease module) of m nodes undergoes a phase transition under certain conditions (*34*, *35*): If p is above p_c^m, some fraction of nodes continue to form an observable module; if, however, p is below p_c^m, the module becomes too fragmented to be observable (Fig. 1E; see also fig. S14 and SM section 6). To quantify this phenomenon, we calculated the minimum network coverage p_c^m required to observe a disease module of original size m, finding that $p_c^m \sim 1/m$, valid for an arbitrary degree distribution of the underlying interactome. Figure 1F illustrates a signature of this phenomenon in the interactome: The observable disease module size S_i versus the number of disease genes associated with each disease follows the predicted percolation transition (purple line). Hence, percolation theory predicts that for diseases with fewer than $N_c \approx 25$ genes, the module is too fragmented to be observable in the current interactome; only diseases with $N_d > N_c$ disease genes should have an observable disease module.

To test whether the observed disease modules represent nonrandom disease gene aggregations, for each disease we compared the size S_i of its observable module with the expected S_i^{rand} if the same number of disease proteins were placed randomly on the interactome. For example, for multiple sclerosis, the observed $S_i = 11$ is significantly larger than the random expectation $S_i^{\text{rand}} = 2 \pm 1$ (z score = 5.8, p value = 3.3×10^{-9}, Figs. 1D and 2A); hence, the observed multiple sclerosis module cannot be attributed to a random agglomeration of disease genes. We also determined for each disease protein the network-based distance d_s to the closest other protein associated with the same disease. Again, for multiple sclerosis, $P(d_s)$ is shifted toward smaller d_s compared to the random expectation $P^{\text{rand}}(d_s)$ (p value = 2.6×10^{-6}, Fig. 2B), indicating that the disconnected disease proteins agglomerate in the neighborhood of the observable module. Altogether, disease genes associated with 226

[1] Center for Complex Networks Research and Department of Physics, Northeastern University, 110 Forsyth Street, 111 Dana Research Center, Boston, MA 02115, USA. [2] Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02215, USA. [3] Center for Network Science, Central European University, Nador u. 9, 1051 Budapest, Hungary. [4] Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA. [5] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA.
*Corresponding author. E-mail: alb@neu.edu

of the 299 diseases show a statistically significant tendency to form disease modules based on both S_i and $P(d_s)$ (fig. S4).

We also asked if there is a relationship between the tendency of disease proteins to agglomerate in the same interactome neighborhood and their biological similarity (7, 36, 37). We find that as the relative size $s_i \equiv S_i/N_i$ of the observable module increases from 0.1 to 0.8, a sign

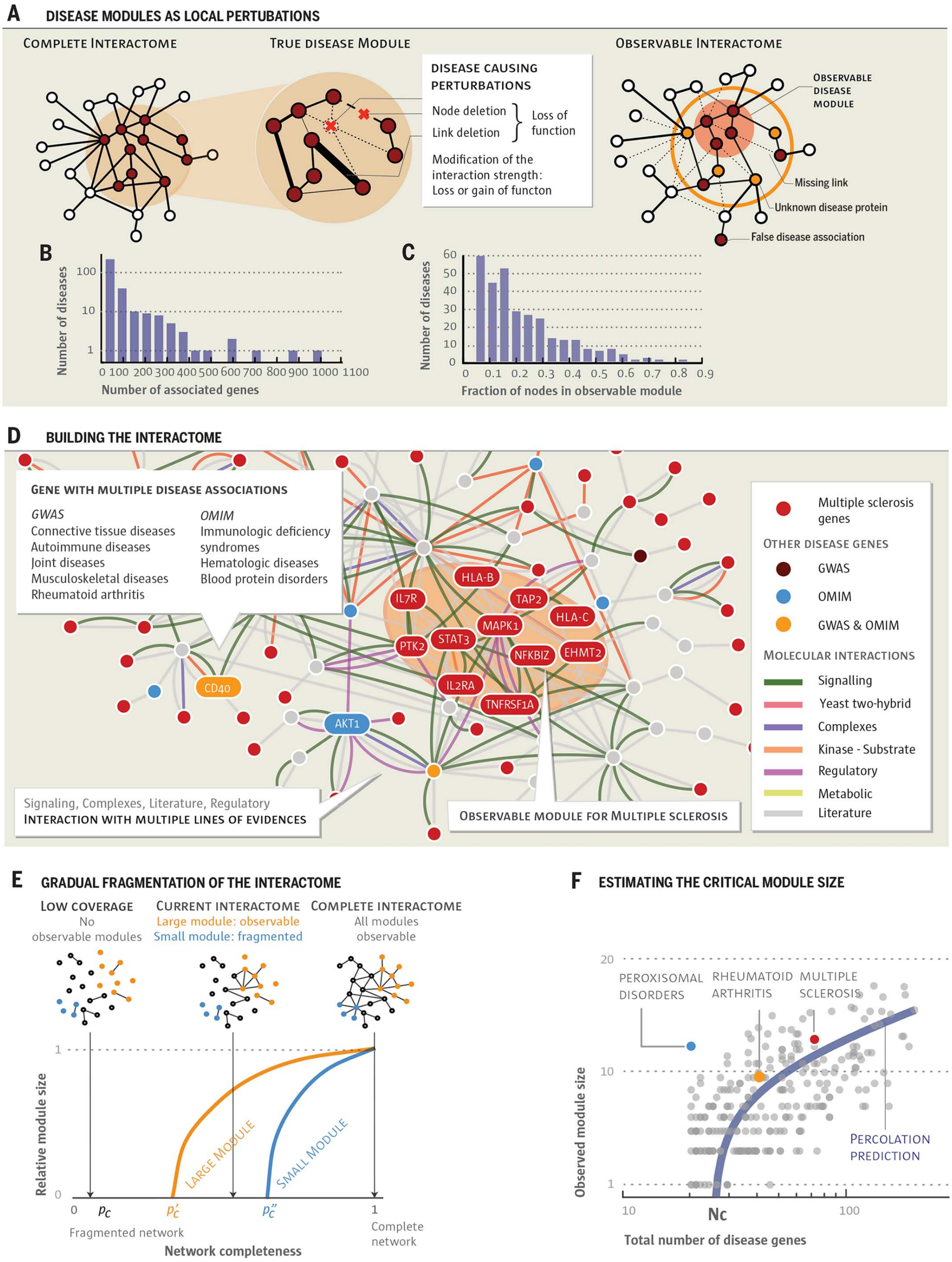


Fig. 1. From the human interactome to disease modules. (**A**) According to the disease module hypothesis, a disease represents a local perturbation of the underlying disease-associated subgraph. Such perturbations could represent the removal of a protein (e.g., by a nonsense mutation), the disruption of a protein-protein interaction, or modifications in the strength of an interaction. The complete disease module can be identified only in a full interactome map; the disease module observable to us captures a subset of this module, owing to data incompleteness. (**B**) Distribution of the number of disease-associated genes for 299 diseases. (**C**) Distribution of the fraction of disease genes within the observable disease module. (**D**) A small neighborhood of the interactome showing the biological nature of each physical interaction and the origin of the disease-gene associations used in our study (see also SM section 1). Genes associated with multiple sclerosis are shown in red, the shaded area indicating their observable module, a connected subgraph consisting of 11 proteins. (**E**) Schematic illustration of the predicted size of the observable disease modules (subgraphs) as a function of network completeness. Large modules should be observable even for low network coverage; to discover smaller modules, we need higher network completeness. (**F**) Size of the observable module as a function of the total number of disease genes. The purple curve corresponds to the percolation-based prediction (SM section 6), indicating that diseases with $N_d < N_c \approx 25$ genes do not have an observable disease module in the current interactome. Each gray point captures one of the 299 diseases.

of increasing agglomeration of the disease genes, the significance of the biological similarity in Gene Ontology (GO) annotations (biological processes, molecular function, and cellular component) increases 10- to 100-fold (Fig. 2, C to E, and fig. S3, a to c), an exceptionally strong effect (see SM sect. 2 for statistical analysis). Similarly, as the mean shortest distance between disease proteins increases from 1 (agglomerated disease proteins) to 3 (scattered disease proteins), we observe a factor of 10 to 100 decrease in the significance of GO term similarity (Fig. 2, F to H, and fig. S3, d to f).

Taken together, we find that genes associated with the same disease tend to agglomerate in the same neighborhood of the interactome. Indeed, although ~80% of the disease proteins are disconnected from the observable module, these isolates tend to be localized in its network vicinity. This result offers quantitative support to the hypothesis that many local neighborhoods of the interactome represent the observable parts of the true, larger and denser disease modules.

Relationship between diseases

If two disease modules overlap, local perturbations leading to one disease will likely disrupt pathways involved in the other disease module as well, resulting in shared clinical characteristics. To test the validity of this hypothesis, we introduce the network-based separation of a disease pair, A and B (Fig. 3A; see also figs. S5 to S7) using

$$s_{AB} \equiv \langle d_{AB} \rangle - \frac{\langle d_{AA} \rangle + \langle d_{BB} \rangle}{2} \qquad (1)$$

s_{AB} compares the shortest distances between proteins within each disease, $\langle d_{AA} \rangle$ and $\langle d_{BB} \rangle$, to the shortest distances $\langle d_{AB} \rangle$ between A-B protein pairs. Proteins associated with both A and B have $d_{AB} = 0$. As discussed in SM section 3.3, the generalization of s_{AB} to account for directed regulatory and signaling interactions does not alter our subsequent findings (fig. S8).

We find that only 7% of disease pairs have overlapping disease neighborhoods with negative s_{AB} (Fig. 3B); the remaining 93% have a positive s_{AB}, indicating that their disease modules are topologically separated (Fig. 3C). Because we lack unambiguous true positive and true negative disease relationships that could be used as a reference, we use two complementary null models to evaluate the statistical significance of each disease pair compared to random expectation (see SM section 2.2). At a global false discovery level of 5%, we find that 75% of all disease pairs exhibit significant s_{AB}. To determine the degree to which this network-based separation of two diseases is predictive for pathobiological manifestations, we rely on four data sets:

1) Biological similarity: We find that the closer two diseases are in the interactome, the higher the GO annotation–based similarity of the proteins associated with them (Fig. 3, D to F). The effect is strong, resulting in a two-order-of-magnitude decrease in GO term similarity as we move from highly overlapping ($s_{AB} \approx -2$) to well-separated disease pairs ($s_{AB} > 0$).

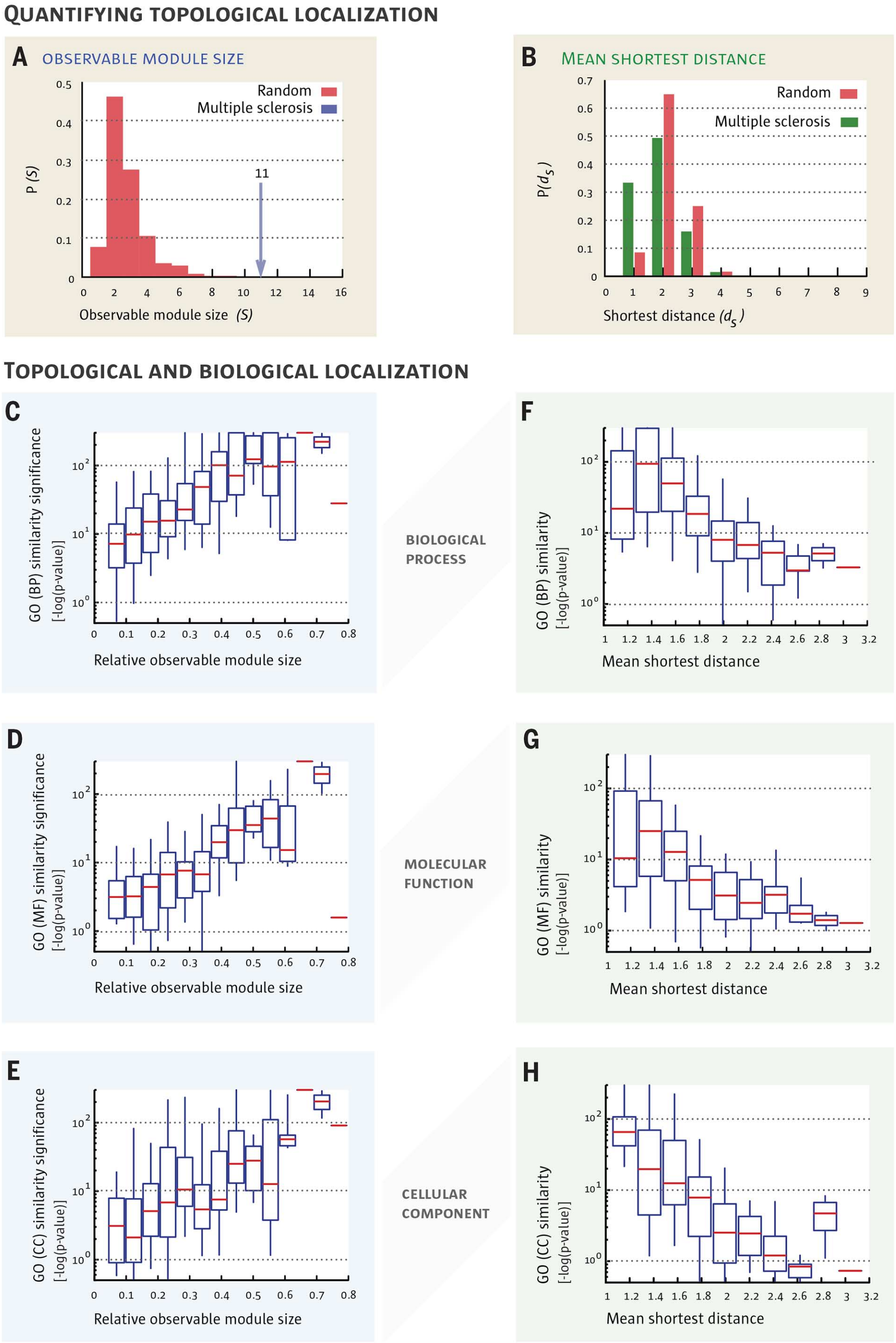


Fig. 2. Topological localization and biological similarity of disease genes. (**A**) The size of the largest connected component S of proteins associated with the same disease shown for multiple sclerosis. The observed module size, S = 11, is significantly larger than the random expectation S^{rand} = 2 ± 1. (**B**) The distribution of the shortest distance of each disease protein to the next closest disease protein d_s. For multiple sclerosis, $P(d_s)$ is significantly shifted compared to the random expectation, indicating that disease genes tend to agglomerate in each other's network neighborhood. (**C** to **H**) The degree of the network-based localization of a disease, as measured by the relative size of its observable module $s_i = S_i/N_d$ and the mean shortest distance $\langle d_s \rangle$, correlates strongly with the significance of the biological similarity of the respective disease genes. Using the GO annotations, we determine for each disease how similar its associated genes are in terms of their biological processes (C and F), molecular function (D and G), and cellular component (E and H). Comparing the resulting values with random expectation, we find that the more localized a disease is topologically (i.e., the larger s_i or the shorter $\langle d_s \rangle$), the higher the significance in the similarity of the associated genes.

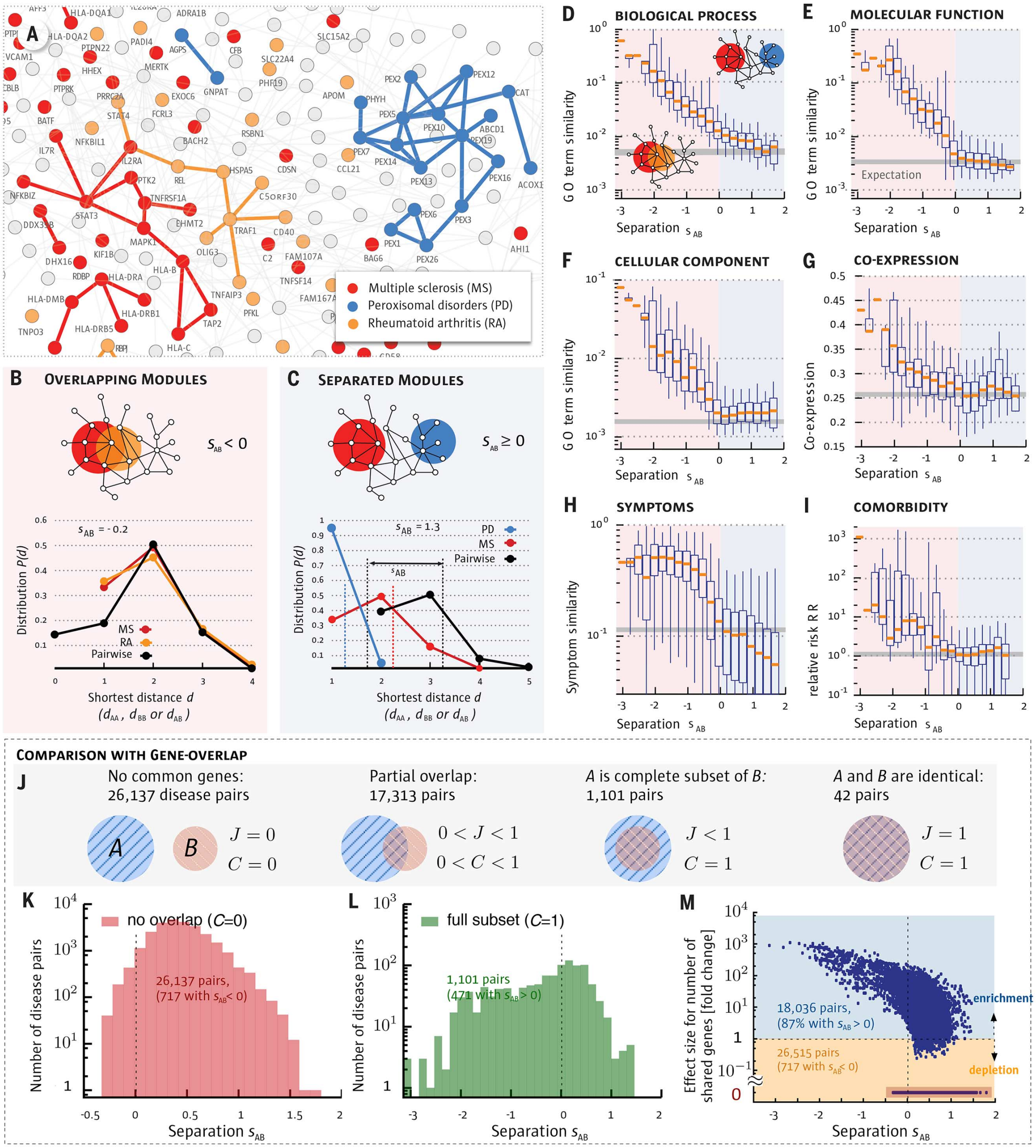


Fig. 3. Network separation and disease similarity. (**A**) A subnetwork of the full interactome highlighting the network-based relationship between disease genes associated with three diseases identified in the legend. (**B** and **C**) Distance distributions for disease pairs that have topologically overlapping modules ($s_{AB} < 0$, B) or topologically separated modules ($s_{AB} > 0$, C). The plots show $P(d)$ for the disease pairs shown in (A). (**D** to **I**) Topological separation versus biomedical similarity. (D to F) GO term similarity; (G) gene coexpression; (H) symptom similarity for all disease pairs in function of their topological separation s_{AB}. The region of overlapping disease pairs is highlighted in red ($s_{AB} < 0$); the region of the separated disease pairs is shown in blue ($s_{AB} > 0$). For symptom similarity, we show the cosine similarity ($c_{AB} = 0$ if there are no shared symptoms between diseases A and B and $c_{AB} = 1$ for diseases with identical symptoms). Comorbidity in (I) is measured by the relative risk *RR* (40). Bars in (D) to (I) indicate random expectation (SM section 1): in (D) to (G), the expected value for a randomly chosen protein pair is shown. In (H) and (I), the mean value of all disease pairs is used. (**J** to **M**) The interplay between gene-set overlap and the network-based relationships between disease pairs. (J) The relationship between gene sets *A* and *B* is captured by the overlap coefficient $C = |A \cap B|/\min(|A|, |B|)$ and the Jaccard-index $J = |A \cap B|/|A \cup B|$. More than half (59%) of the disease pairs do not share genes ($J = C = 0$); hence, their relation cannot be uncovered based on shared genes. (K) Distribution of s_{AB} for disease pairs with no gene overlap. We find that despite having disjoint gene sets, 717 diseases pairs have overlapping modules ($s_{AB} < 0$). (L) The distribution of s_{AB} for disease pairs with complete gene overlap ($C = 1$) shows a broad range of network-based relationships, including non-overlapping modules ($s_{AB} > 0$). (M) Fold change of the number of shared genes compared to random expectation versus s_{AB} for all disease pairs. The 59% of all disease pairs without shared genes are highlighted with red background. For 98% of all disease pairs that share at least one gene, the gene-based overlap is larger than expected by chance. Nevertheless, most (87%) of these disease pairs are separated in the network ($s_{AB} > 0$). Conversely, a considerable number of pairs (717) without shared genes exhibit detectable network overlap ($s_{AB} < 0$).

2) Coexpression: We find that the coexpression-based correlation across 70 tissues ([36]) between genes associated with overlapping diseases is almost twice that of well-separated diseases (Fig. 3G), falling to the random expectation for $s_{AB} > 0$.

3) Disease symptoms: We find that symptom similarity, as captured by large-scale medical bibliographic records ([38]), falls about an order of magnitude as we move from overlapping ($s_{AB} < 0$) to separated ($s_{AB} > 0$) diseases (Fig. 3H). Non-overlapping diseases share fewer symptoms than expected by chance.

4) Comorbidity: We used the disease history of 30 million individuals aged 65 and older (U.S. Medicare) to determine for each disease pair the relative risk RR of disease comorbidity ([39]) (Fig. 3I), finding that the relative risk drops from $RR \geq 10$ for $s_{AB} < 0$ to the random expectation of $RR \approx 1$ for $s_{AB} > 0$.

Thus, the network-based distance of two diseases indicates their pathobiological and clinical similarity. This result suggests a molecular network model of human disease: Each disease has a well-defined location and a diameter $\langle d_{AA} \rangle$ that captures its network-based size (Fig. 3, A to C). If two disease modules are topologically separated ($s_{AB} > 0$), then the diseases are pathobiologically distinct. If the disease modules topologically overlap ($s_{AB} < 0$), the magnitude of the overlap is indicative of their biological relationship: The higher the overlap, the more significant are the pathobiological similarities between them. We, therefore, represent each disease by a sphere with diameter $\langle d_{AA} \rangle$ in a three-dimensional (3D) disease space such that the physical distance r_{AB} between diseases A and B correlates with the observed network-based distance $\langle d_{AB} \rangle$ (Fig. 4A; see also fig. S15 and SM section 8). Disease modules that do not overlap in Fig. 4A are predicted to be pathobiologically distinct; for those that overlap, the degree of overlap captures their common pathobiology and phenotypic characteristics.

To test the predictive power of this model, we grouped the disease pairs with $s_{AB} < 0$ into the "overlapping" disease category, and those with $s_{AB} > 0$ into the "non-overlapping" disease category. As Fig. 4, B to G, indicates, all biological and clinical characteristics show statistically highly significant similarity for overlapping diseases, whereas the effects vanish for the non-overlapping disease pairs.

The disease separation allows us to identify unexpected overlapping disease pairs, i.e., those that lack overt pathobiological or clinical association (see table S1 for 12 such examples). For example, we find that asthma, a respiratory disease, and celiac disease, an autoimmune disease of the small intestine, are localized in overlapping neighborhoods ($s_{AB} < 0$, Fig. 4N), suggesting shared molecular roots despite their rather different pathobiologies. A closer inspection reveals evidence supporting this prediction: The two diseases share three genes identified via genome-wide associations with genome-wide significance (*HLA-DQA1*, *IL18R1*, *IL1RL1*), and, recently, SNP rs1464510, previously associated with celiac disease, was also found to be associated with asthma ([40]). Although the two diseases have few common phenotypic features, they exhibit a remarkably high comorbidity (RR = 6.18) and statistically significant coexpression between their genes (r = 0.32, p value = 0.02). Furthermore, the top enriched pathway in the combined gene set of the two diseases is the immune network for immunoglobulin A (IgA) production (p value = 5×10^{-15}, Fig. 4O) with 48 genes, of which seven are associated with asthma and five with celiac disease. Measuring amounts of an IgA antibody subclass against tissue transglutaminase (ATA) is widely used to screen for and diagnose celiac disease ([41]). At the same time, the IgA response to allergens in the respiratory tract of asthma patients plays a pathogenic role through eosinophil activation ([42]).

To determine whether we could have arrived at the same conclusion by identifying diseases with shared genes ([7]), we quantified the predictive power of gene overlap, finding that, indeed, disease pairs with large gene overlap tend to be localized in the same network neighborhood (Fig. 3, L and M). Nevertheless, 59% of disease pairs do not share genes; hence, their relationship cannot be resolved based on the shared gene hypothesis (Fig. 3J; see also figs. S9 and S10). We, therefore, repeated the analysis of Fig. 4, B to G, for all disease pairs without common genes, finding that s_{AB} continues to predict accurately the biological similarity (or distinctness) of these disease pairs (Fig. 4, H to M, and SM section 3). Overall, we find 717 pairs with overlapping disease modules ($s_{AB} < 0$, Fig. 3K), relationships that cannot be predicted based on gene overlap. For example, lymphoma, a cancer, and myocardial infarction, a heart disease, do not share disease genes. Yet, they have strongly overlapping modules ($s_{AB} = -0.24$), indicating that they are located in the same neighborhood of the interactome. Indeed, we find that SMARCA4, a protein associated with myocardial infarction, interacts with ALK, MYC, and NF-κB2, which are lymphoma disease proteins. Cancer cells frequently depend on chromatin regulatory activities to maintain a malignant phenotype. It has been shown that leukemia cells require the SWI/SNF chromatin remodeling complex containing the SMARCA4 protein as the catalytic subunit for their survival and aberrant self-renewal potential ([43]). The relatedness of the two diseases is further supported by a high comorbidity [relative risk (RR) = 2.1] and the clinical finding that intravascular large cell lymphoma can affect and obstruct the small vessels of the heart ([44]). Other disease pairs that lack shared genes but are found in the same neighborhood of the interactome include glioma and gout, glioma and myocardial infarction, and myeloproliferative disorders and proteinuria, each pair having high comorbidity (RR = 2.43, 6.3, and 2.0, respectively). A detailed discussion of these and other novel disease-disease relationships predicted by our approach is offered in SM section 10.

Summary and discussion

A complete and accurate map of the interactome could have tremendous impact on our ability to understand the molecular underpinnings of human disease. Yet, such a map is at least a decade away, which makes it currently impossible to evaluate precisely how far a given disease module is from completion. Yet, here we showed that despite its incompleteness, the available interactome has sufficient coverage to pursue a systematic network-based approach to human diseases. To be specific, we offer quantitative evidence for the identifiability of some disease modules, while showing that for other diseases the identifiability condition is not yet satisfied at the current level of incompleteness of the interactome. Most important, we demonstrated that the relative interactome-based position of two disease modules is a strong predictor of their biological and phenotypic similarity. Throughout this paper, we focused on the impact of network incompleteness, ignoring another limitation of the interactome: It is prone to notable investigative biases ([12], [32], [33]) (see also fig. S13 and SM section 5). We, therefore, repeated our analysis relying only on high-throughput data from yeast two-hybrid screens ([12]) (y2h, SM section 4), finding that the diameter $\langle d_{AA} \rangle$ of the observable modules, the distance $\langle d_{AB} \rangle$ and separation s_{AB} of all disease pairs measured in the full and the unbiased interactome show statistically highly significant correlations. Similarly, OMIM is also prone to selection and investigative biases; hence, we repeated our measurements using only unbiased GWAS-associated disease genes. Comparing gene sets that include OMIM data and those that only contain GWAS associations, we again find highly significant correlations for $\langle d_{AA} \rangle$, $\langle d_{AB} \rangle$, and s_{AB} (figs. S11 and S12). Therefore, the disease modules and the overlap between them can be reproduced in the unbiased data as well, indicating that our key findings cannot be attributed to investigative biases. We estimate the minimal number of associated genes that a disease needs to have in order to be observable to be around 25 for the current interactome. Unbiased high-throughput data alone have not yet reached sufficient coverage to map out putative modules for many diseases; For the y2h network, being a subset of the interactome with a much lower coverage, the respective minimal number is around 350 (N_c^{y2h}); hence, only a few disease modules can be observed (see fig. S14f). However, this approach can provide valuable insights into the properties of the complete interactome (SM section 6). Indeed, as the current y2h data are expected to represent a uniform subset of the complete y2h network ([12]), we can use it to derive the minimum coverage p_c^m of the latter. As the coverage of high-throughput maps improves, they will allow us to use the full power of unbiased approaches for disease module identification.

The true value of the developed interactome-based approach is its open-ended multipurpose nature: It offers a platform that can address numerous fundamental and practical issues pertaining to our understanding of human disease. This platform can be used to improve the interpretation of GWAS data (see fig. S16 and SM section 10 for an application to type II

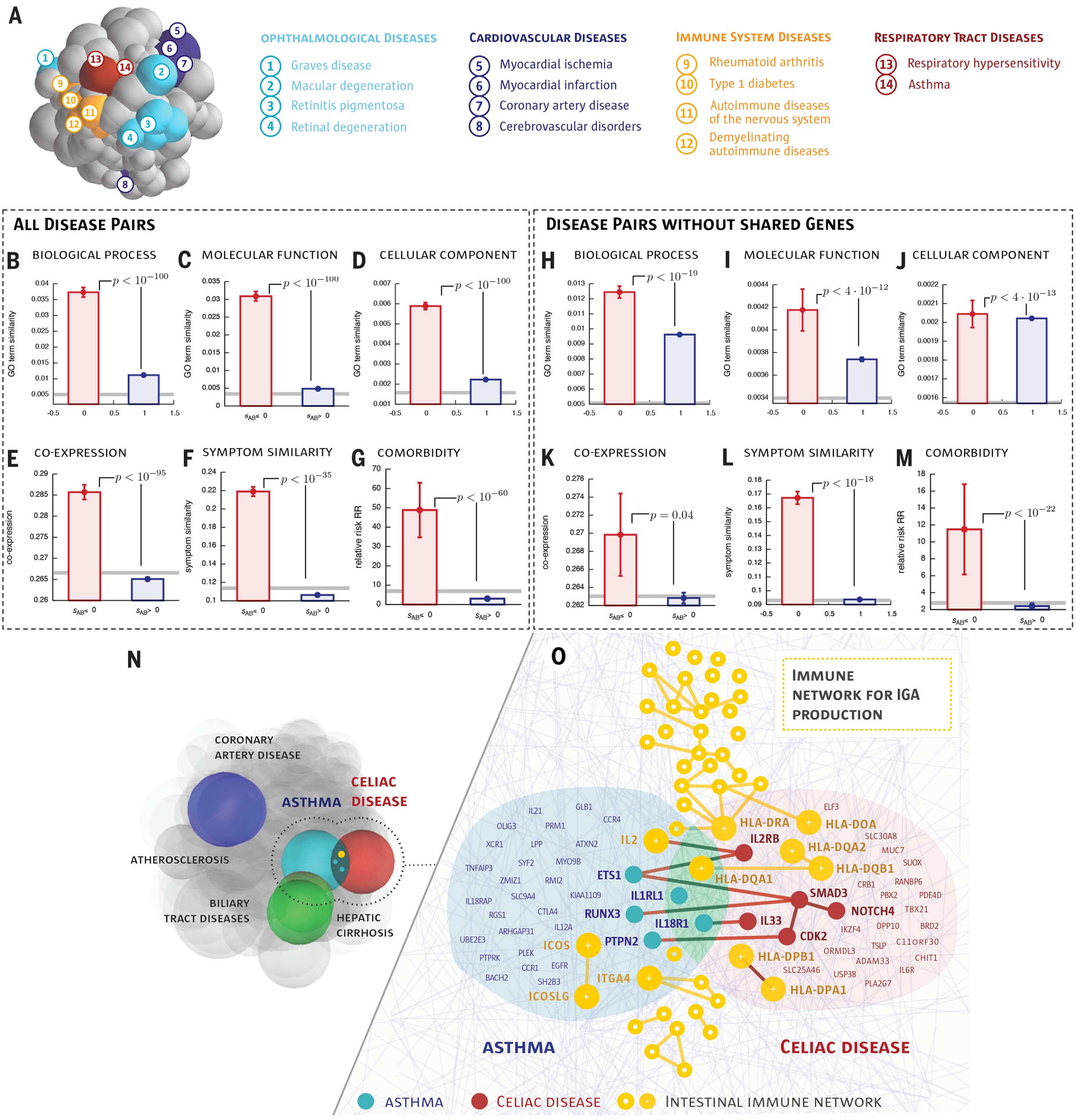


Fig. 4. Network-based model of disease-disease relationship. (**A**) To illustrate the uncovered network-based relationship between diseases, we place each disease in a 3D disease space, such that their physical distance to other diseases is proportional to $\langle d_{AB} \rangle$ predicted by the interactome-based analysis. Diseases whose modules (spheres) overlap are predicted to have common molecular underpinnings. The colors capture several broad disease classes, indicating that typically diseases of the same class are located close to each other. There are exceptions, such as cerebrovascular disease, which is separated from other cardiovascular diseases, suggesting distinct molecular roots. (**B** to **G**) Biological similarity shown separately for the predicted overlapping and non-overlapping disease pairs (see Fig. 3, D to I, for interpretation). Error bars indicate the SEM. Gray lines show random expectation, either for random protein pairs (B to E, H to K) or for a random disease pair (F, G, L, M); *p* values denote the significance of the difference of the means according to a Mann-Whitney U test. (**H** to **M**) Biological similarity for disease pairs that do not share genes (control set). (**N**) Three overlapping disease pairs in the disease space. Coronary artery diseases and atherosclerosis, as well as hepatic cirrhosis and biliary tract diseases, are diseases with common classification; hence, their disease modules overlap. Our methodology also predicts several overlapping disease modules of apparently unrelated disease pairs (table S1), illustrated by asthma and celiac disease. (**O**) A network-level map of the overlapping asthma–celiac disease network neighborhood; also shown is the IgA production pathway (yellow) that plays a biological role in both diseases. We denote genes that are either shared by the two diseases or by the pathway, or that interact across the modules.

diabetes), help us uncover new uses for existing drugs (repurposing) by identifying the disease modules located in the vicinity of each drug target (45–47), and facilitate the discovery of the molecular underpinnings of undiagnosed diseases by exploiting the agglomeration of mutations and expression changes in network neighborhoods associated with well-characterized diseases. In the long run, network-based approaches, relying on an increasingly accurate interactome, are poised to become highly useful in interpreting disease-associated genome variations.

Materials and methods

Interactome construction

We combine several sources of protein interactions: (i) regulatory interactions derived from transcription factors binding to regulatory elements; (ii) binary interactions from several yeast two-hybrid high-throughput and literature-curated data sets; (iii) literature-curated interactions derived mostly from low-throughput experiments; (iv) metabolic enzyme-coupled interactions; (v) protein complexes; (vi) kinase-substrate pairs; and (vii) signaling interactions. The union of all interactions from (i) to (vii) yields a network of 13,460 proteins that are interconnected by 141,296 interactions. For more information on the individual data sets and general properties of the interactome, see SM section 1.

Disease-gene associations

We integrate disease-gene annotations from Online Mendelian Inheritance in Man (OMIM; www.ncbi.nlm.nih.gov/omim) (48) and UniProtKB/Swiss-Prot as compiled by (30) with GWAS data from the Phenotype-Genotype Integrator database (PheGenI; www.ncbi.nlm.nih.gov/gap/PheGenI) (31), using a genome-wide significance cutoff of p value $\leq 5 \times 10^{-8}$. To combine the different disease nomenclatures of the two sources into a single standard vocabulary, we use the Medical Subject Headings ontology (MeSH; www.nlm.nih.gov/mesh/) as described in SM section 1. After filtering for diseases with at least 20 associated genes and genes for which we have interaction information, we obtain 299 diseases and 3173 associated genes.

Additional disease and gene annotation data

For the analysis of the similarity between genes and diseases, we use (i) Gene Ontology (GO) annotations (49); (ii) tissue-specific gene expression data (36); (iii) symptom disease associations (38); (iv) comorbidity data (39); and (v) pathway annotations from the Molecular Signatures Database (MSigDB) (50). Full details on data sources, processing, and analysis are provided in SM section 1.

Network localization

We use two complementary measures to quantify the degree to which disease proteins agglomerate in specific interactome neighborhoods: (i) observable module size S, representing the size of the largest connected subgraph formed by disease proteins; and (ii) shortest distance d_s. For each of the N_d disease proteins, we determine the distance d_s to the next-closest protein associated with the same disease. The average $\langle d_s \rangle$ can be interpreted as the diameter of a disease on the interactome. The network-based overlap between two diseases A and B is measured by comparing the diameters $\langle d_{AA} \rangle$ and $\langle d_{BB} \rangle$ of the respective diseases to the mean shortest distance $\langle d_{AB} \rangle$ between their proteins: $s_{AB} = \langle d_{AB} \rangle - (\langle d_{AA} \rangle + \langle d_{BB} \rangle)/2$. Positive s_{AB} indicates that the two disease modules are separated on the interactome, whereas negative values correspond to overlapping modules. Details on the analysis and the appropriate random controls are presented in SM section 2.

Gene-based disease overlap

The overlap between two gene sets A and B is measured by the overlap coefficient $C = |A \cap B| / \min(|A|,|B|)$ and the Jaccard-index $J = |A \cap B| / |A \cup B|$. The values of both measures lie in the range [0,1] with $J,C = 0$ for no common genes. A Jaccard-index $J = 1$ indicates two identical gene sets, whereas the overlap coefficient $C = 1$ when one set is a complete subset of the other. For a statistical evaluation of the observed overlaps, we use a basic hypergeometric model with the null hypothesis that disease-associated genes are randomly drawn from the space of all N genes in the network (see SM section 3 for full details).

REFERENCES AND NOTES

1. M. Buchanan, G. Caldarelli, P. De Los Rios, *Networks in Cell Biology* (Cambridge Univ. Press, Cambridge, 2010).
2. T. Pawson, R. Linding, Network medicine. *FEBS Lett.* **582**, 1266–1270 (2008). doi: 10.1016/j.febslet.2008.02.011; pmid: 18282479
3. E. E. Schadt, Molecular networks as sensors and drivers of common human diseases. *Nature* **461**, 218–223 (2009). doi: 10.1038/nature08454; pmid: 19741703
4. A. Califano, A. J. Butte, S. Friend, T. Ideker, E. Schadt, Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* **44**, 841–847 (2012). doi: 10.1038/ng.2355; pmid: 22836096
5. A. Zanzoni, M. Soler-López, P. Aloy, A network medicine approach to human disease. *FEBS Lett.* **583**, 1759–1765 (2009). doi: 10.1016/j.febslet.2009.03.001; pmid: 19269289
6. A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011). doi: 10.1038/nrg2918; pmid: 21164525
7. K.-I. Goh *et al*., The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 8685–8690 (2007). doi: 10.1073/pnas.0701361104; pmid: 17502601
8. M. Oti, B. Snel, M. A. Huynen, H. G. Brunner, Predicting disease genes using protein-protein interactions. *J. Med. Genet.* **43**, 691–698 (2006). doi: 10.1136/jmg.2006.041376; pmid: 16611749
9. K. Lage *et al*., Dissecting spatio-temporal protein networks driving human heart development and related disorders. *Mol. Syst. Biol.* **6**, 381 (2010). doi: 10.1038/msb.2010.36; pmid: 20571530
10. H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007). doi: 10.1038/msb4100180; pmid: 17940530
11. R. Mosca, T. Pons, A. Céol, A. Valencia, P. Aloy, Towards a detailed atlas of protein-protein interactions. *Curr. Opin. Struct. Biol.* **23**, 929–940 (2013). doi: 10.1016/j.sbi.2013.07.005; pmid: 23896349
12. T. Rolland *et al*., A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014). doi: 10.1016/j.cell.2014.10.050; pmid: 25416956
13. G. T. Hart, A. K. Ramani, E. M. Marcotte, How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7**, 120 (2006). doi: 10.1186/gb-2006-7-11-120; pmid: 17147767
14. K. Venkatesan *et al*., An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009). pmid: 19060904
15. M. P. Stumpf *et al*., Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 6959–6964 (2008). doi: 10.1073/pnas.0708078105; pmid: 18474861
16. M. N. Wass, A. David, M. J. Sternberg, Challenges for the prediction of macromolecular interactions. *Curr. Opin. Struct. Biol.* **21**, 382–390 (2011). doi: 10.1016/j.sbi.2011.03.013; pmid: 21497504
17. J. Xu, Y. Li, Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* **22**, 2800–2805 (2006). doi: 10.1093/bioinformatics/btl467; pmid: 16954137
18. I. Feldman, A. Rzhetsky, D. Vitkup, Network properties of genes harboring inherited disease mutations. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4323–4328 (2008). doi: 10.1073/pnas.0701722105; pmid: 18326631
19. M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, A. Rzhetsky, Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15148–15153 (2004). doi: 10.1073/pnas.0404315101; pmid: 15471992
20. L. Franke *et al*., Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* **78**, 1011–1025 (2006). doi: 10.1086/504300; pmid: 16685651
21. S. Köhler, S. Bauer, D. Horn, P. N. Robinson, Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008). doi: 10.1016/j.ajhg.2008.02.013; pmid: 18371930
22. Y. Chen *et al*., Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008). doi: 10.1038/nature06757; pmid: 18344982
23. S. E. Baranzini *et al*., Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.* **18**, 2078–2090 (2009). doi: 10.1093/hmg/ddp120; pmid: 19286671
24. C. E. Wheelock *et al*., Systems biology approaches and pathway tools for investigating cardiovascular disease. *Mol. Biosyst.* **5**, 588–602 (2009). doi: 10.1039/b902356a; pmid: 19462016
25. A. S. Khalil, J. J. Collins, Synthetic biology: Applications come of age. *Nat. Rev. Genet.* **11**, 367–379 (2010). doi: 10.1038/nrg2775; pmid: 20395970
26. S. Wuchty *et al*., Gene pathways and subnetworks distinguish between major glioma subtypes and elucidate potential underlying biology. *J. Biomed. Inform.* **43**, 945–952 (2010). doi: 10.1016/j.jbi.2010.08.011; pmid: 20828632
27. I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, E. M. Marcotte, Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011). doi: 10.1101/gr.118992.110; pmid: 21536720
28. U. M. Singh-Blom *et al*., Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLOS ONE* **8**, e58977 (2013). doi: 10.1371/journal.pone.0058977; pmid: 23650495
29. A. Rzhetsky, D. Wajngurt, N. Park, T. Zheng, Probing genetic overlap among complex human phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 11694–11699 (2007). doi: 10.1073/pnas.0704820104; pmid: 17609372
30. A. Mottaz, Y. L. Yip, P. Ruch, A.-L. Veuthey, Mapping proteins to disease terminologies: From UniProt to MeSH. *BMC Bioinformatics* **9** (suppl. 5), S3 (2008). doi: 10.1186/1471-2105-9-S5-S3; pmid: 18460185
31. E. M. Ramos *et al*., Phenotype-Genotype Integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.* **22**, 144–147 (2014). doi: 10.1038/ejhg.2013.96; pmid: 23695286
32. L. Hakes, J. W. Pinney, D. L. Robertson, S. C. Lovell, Protein-protein interaction networks and biology—What's the connection? *Nat. Biotechnol.* **26**, 69–72 (2008). doi: 10.1038/nbt0108-69
33. M. E. Cusick *et al*., Literature-curated protein interaction datasets. *Nat. Methods* **6**, 39–46 (2009). doi: 10.1038/nmeth.1284; pmid: 19116613
34. R. Cohen, S. Havlin, *Complex Networks. Structure, Robustness and Function* (Cambridge Univ., Cambridge, 2010).
35. S. Bornholdt, H. G. Schuster, Eds., *Handbook of Graphs and Networks* (Wiley Online Library, 2003), vol. 2.
36. A. I. Su *et al*., A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 6062–6067 (2004). doi: 10.1073/pnas.0400782101; pmid: 15075390
37. T. K. Gandhi *et al*., Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* **38**, 285–293 (2006). pmid: 16501559

38. X. Zhou, J. Menche, A.-L. Barabási, A. Sharma, Human symptoms-disease network. *Nat. Commun.* **5**, 4212 (2014). doi: 10.1038/ncomms5212; pmid: 24967666
39. C. A. Hidalgo, N. Blumm, A.-L. Barabási, N. A. Christakis, A dynamic network approach for the study of human phenotypes. *PLOS Comput. Biol.* **5**, e1000353 (2009). doi: 10.1371/journal.pcbi.1000353; pmid: 19360091
40. K. A. Hunt *et al*., Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* **40**, 395–402 (2008). doi: 10.1038/ng.102; pmid: 18311140
41. D. A. van der Windt, P. Jellema, C. J. Mulder, C. M. Kneepkens, H. E. van der Horst, Diagnostic testing for celiac disease among patients with abdominal symptoms: A systematic review. *JAMA* **303**, 1738–1746 (2010). doi: 10.1001/jama.2010.549; pmid: 20442390
42. C. Pilette, S. R. Durham, J.-P. Vaerman, Y. Sibille, Mucosal immunity in asthma and chronic obstructive pulmonary disease: A role for immunoglobulin A? *Proc. Am. Thorac. Soc.* **1**, 125–135 (2004). doi: 10.1513/pats.2306032
43. J. Shi *et al*., Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev.* **27**, 2648–2662 (2013). doi: 10.1101/gad.232710.113; pmid: 24285714
44. A. Bauer, B. Perras, S. Sufke, H.-P. Horny, B. Kreft, Myocardial infarction as an uncommon clinical manifestation of intravascular large cell lymphoma. *Acta Cardiol.* **60**, 551–555 (2005). doi: 10.2143/AC.60.5.2004979; pmid: 16261789
45. A. L. Hopkins, Network pharmacology: The next paradigm in drug discovery. *Nat. Chem. Biol.* **4**, 682–690 (2008). doi: 10.1038/nchembio.118; pmid: 18936753
46. J. Mestres, E. Gregori-Puigjané, S. Valverde, R. V. Solé, The topology of drug-target interaction networks: Implicit dependence on drug properties and target families. *Mol. Biosyst.* **5**, 1051–1057 (2009). doi: 10.1039/b905821b; pmid: 19668871
47. M. Kuhn *et al*., Systematic identification of proteins that elicit drug side effects. *Mol. Syst. Biol.* **9**, 663 (2013). doi: 10.1038/msb.2013.10; pmid: 23632385
48. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005). doi: 10.1093/nar/gki033; pmid: 15608251
49. M. Ashburner *et al*.The Gene Ontology Consortium, Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000). doi: 10.1038/75556; pmid: 10802651
50. A. Subramanian *et al*., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005). doi: 10.1073/pnas.0506580102 pmid: 16199517

ACKNOWLEDGMENTS

We thank A.-R. Carvunis, S. Pevzner, and T. Rolland for providing invaluable insights into the y2h data set; J. Bagrow and F. Simini for many discussions on the network methods; and G. Musella for figure design. This work was supported by NIH grants P50-HG004233, U01-HG001715, and U01-HG007690 from NHGRI and P01-HL083069, R37-HL061795, RC2-HL101543, and U01-HL108630 from NHLBI.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/347/6224/1257601/suppl/DC1
Materials and Methods
Figs. S1 to S16
Table S1
External Databases S1 to S4
Source Code
References (*51*–*107*)

17 June 2014; accepted 15 January 2015
10.1126/science.1257601